Private Emotions vs. Social Interaction – towards New Dimensions in Research on Emotion

Anton Batliner, Stefan Steidl, Christian Hacker, Elmar Nöth, and Heinrich Niemann
Lehrstuhl für Mustererkennung, Martensstr. 3, 91058 Erlangen, F.R. of Germany
batliner@informatik.uni-erlangen.de

ABSTRACT

The 'traditional' first two dimensions in emotion research are AROUSAL and VALENCE. Normally, they are obtained by using elicited, acted data. In this paper, we use realistic, spontaneous speech data from our 'AIBO' corpus (human-robot communication, children interacting with Sony's AIBO robot). The recordings were done in a Wizard-of-Oz scenario: the children believed that AIBO obeys their commands; in fact, AIBO followed a fixed script and often disobeyed. The emotional annotations of five labellers, transformed into a confusion matrix, were used in a non-metrical multi-dimensional scaling to display two dimensions, the first being VALENCE, the second, however, not AROUSAL but INTERACTION, i.e., addressing oneself *(angry, joyful)* or the communication partner *(motherese, reprimanding)*. We show that it depends on the specifity of the scenario and on the subjects' conceptualizations whether this new dimension can be observed, and discuss impacts on the practice of labelling and processing emotional data.

Keywords

emotion, categories, dimensions, annotation, non-metrical multi-dimensional scaling

1. INTRODUCTION

In this introduction, we want to sketch the methodological, theoretical, and historical context of the experiments we will report on.

1.1 Acted vs. Realistic Data

Most of the research on emotion in general and on emotion in speech in particular conducted in the last decades has been on elicited, acted, and by that rather full-blown emotional states. Of course, this means that the data obtained display specific traits: trivially but most importantly, the subjects only displayed those states that they have been told to display. The set of labels is thus **pre-defined**. Secondly, the better actors the subjects were, the more pronounced and by that, easier to tell apart, these emotions were. The models and theories based on such data are normally not called 'data-driven' but based on theoretical considerations. In fact, they are data-driven as well, because they were founded and further developed with the help of these – pre-defined – data. In linguistics and phonetics, the state of affairs had been similar: for decades, tightly controlled (and by that, pre-defined as well) and/or **'interesting'** data were objects of investigation - 'interesting' not because they were representative but because they were distinct and at the same time, well-suited to help deciding between competing theories, models, or explanations. However, when all these models had to be put into real practice, i.e., when real-life, spontaneous speech had to be processed, researchers learned that 'all of a sudden', their data looked pretty much different, and that their models were as such not of much use any longer. In the same vein, in the last decade, non-acted data are considered to be more and more important in research on emotion as well.

1.2 Categories vs. Dimensions

Broadly speaking, there are two different conceptualizations of emotion phenomena that, in practice, are mirrored in the type of annotation performed for databases: the one dates back to W. Wundt [17] and assumes emotional dimensions such as AROUSAL, VALENCE, and CONTROL; emotional phenomena are annotated on continuous scales. Normally, only the first two dimensions are used. In contrast, a discontinuous, categorical conceptualization uses categories like the big *n* emotions (*anger, fear, sadness, disgust,* etc.) which are annotated as such, by using the term that describes best the phenomenon. The two conceptualizations can be mapped onto each other by, e.g., placing category labels onto appropriate positions within the two-dimensional emotional space with AROUSAL as first and VALENCE as second dimension, cf. [8].[1] Normally, this has been achieved by similarity judgment experiments using, e.g., the semantic differential. Here, the precise position in the multidimensional space is obtained empirically; the dimensional terms themselves are pre-defined. The problem might be that these 'traditional' dimensions AROUSAL and VALENCE have been developed by looking at prototypical, acted emotions, be it for speech or for facial gestures. This holds for category labels as well. Matters are different if we go over to real-life data: full-blown emotions are getting less important. As it turns out, interpersonal relations are coming to the fore instead.

1.3 Data and Concepts

A dimension is rather a 'higher level', theoretical concept, encompassing several different categories, and more closely attached to models than categories. The latter ones can, of course, be 'higher level' as well, but they can also be used in pre-theoretical, everyday language. Thus we do not start with pre-defined (acted) data or with pre-defined categorical concepts but we let our subjects decide what they will

[1] In [9] dimensions and categories and their specific advantages and drawbacks are compared.

produce, and we get at our labels via 'inspection' of our data: there were some pilot passes before the final labelling was done, with subsequent exchange between supervisor and labellers.[2]

2. MATERIAL

The general frame for the database reported on in this paper is human-machine – to be more precise, human-robot – communication, children's speech, and the elicitation and subsequent recognition of emotional user states. The robot is the (pet dog-like) Sony's AIBO robot. The basic idea is to combine a new type of corpus (children's speech) with 'natural' emotional speech within a Wizard-of-Oz task. The speech is intended to be 'natural' because children do not disguise their emotions to the same extent as adults do. However, it is of course not fully 'natural' as it might be in a non-supervised setting. Furthermore the speech is spontaneous, because the children were not told to use specific instructions but to talk to the AIBO like they would talk to a friend. In this experimental design, the child is led to believe that the AIBO is responding to his or her commands, but the robot is actually being controlled by a human operator, using the 'AIBO Navigator' software over a wireless LAN (the existing AIBO speech recognition module is not used). The wizard causes the AIBO to perform a fixed, predetermined sequence of actions, which takes no account of what the child says. For the sequence of AIBO's actions, we tried to find a good compromise between obedient and disobedient behaviour: we wanted to provoke the children in order to elicit emotional behaviour but of course we did not want to run the risk that they break off the experiment. The children believed that the AIBO was reacting to their orders - albeit often not immediately. In fact, it was the other way round: the AIBO always strictly followed the same screen-plot, and the children had to align their orders to it's actions. By this means, it is possible to examine different children's reactions to the very same sequence of AIBO's actions. In this paper, we mainly want to deal with the German recordings; the parallel English data recorded at University of Birmingham are described in more detail in [3] and below, in section 4.2. The data were collected from 51 children (age 10 - 13, 21 male, 30 female). The children are from two different schools; the recordings took place in two class-rooms. Each recording session took some 30 minutes. Because of the experimental setup, these recordings contain a huge amount of silence (reaction time of the AIBO), which caused a noticeable reduction of recorded speech after raw segmentation; finally we obtained about 9.2 hours of speech. Based on pause information, the data were segmented automatically into 'utterances'; average number of words per utterance is 3.5.

3. ANNOTATION

The labellers listened to the utterances (no video information was given) of each child in sequential (not randomized) order. Five labellers annotated independently from each other each word as neutral (default) or as belonging to one of ten other classes which were obtained by inspection of the data, cf. above.[3] We do not claim that our labels represent children's emotions in general, only that they are adequate for the modelling of these children's behaviour in this specific scenario. We resort to majority voting (henceforth MV): if three or more labellers agree on the same label, this very label is attributed to the word; if four or five labellers agree, we assume some sort of prototypes. Table 1 shows the labels used and the resp. number # of MV cases for the German and the English data. We will come back to the English figures below, in section 4.2.

Table 1: Emotional labels used with # of majority voting (MV) cases for German and English data

label	# German	# English
joyful	101	11
surprised	0	0
motherese	1261	55
neutral	39177	7171
rest (spurious emotions)	3	0
bored	11	0
emphatic	2528	631
helpless	3	20
touchy (irritated)	225	7
angry	84	23
reprimanding	310	127
no MV	4705	429
total	48408	8474

We consider only labels with more than 50 MVs, resulting in seven classes. *joyful* and *angry* belong to the 'big' emotions, the other ones rather to 'emotion-related/emotion-prone' user states. The state *emphatic* has to be commented on especially: based on our experience with other emotional databases [2], any marked deviation from a neutral speaking style can (but need not) be taken as a possible indication of some (starting) trouble in communication. If a user gets the impression that the machine does not understand her, she tries different strategies – repetitions, re-formulations, other wordings, or simply the use of a pronounced, marked speaking style. Such a style does thus not necessarily indicate any deviation from a neutral user state but it means a higher probability that the (neutral) user state will possibly be changing soon. Of course, it can be something else as well: a user idiosyncrasy, or a special style – 'computer talk' – that some people use while speaking to a computer, like speaking to a non-native, or to a child, or to an elderly person who is hard of hearing. Thus the fact that *emphatic* can be observed can only be interpreted meaningfully if other factors are considered. There are three further – practical – arguments for the annotation of *emphatic*: firstly, it is to a large extent a prosodic phenomenon, thus it can be modelled and classified with prosodic features. Secondly, if the labellers are allowed to label *emphatic* it might be less likely that they confuse it with other user states. Thirdly, we can try and model emphasis as an indication of (arising) problems in the communication[2].

[2]Note that this is of course no 'tabula rasa' approach, but we hope that it is more data-driven and by that, offers more opportunities to detect new phenomena, than a closed approach with a fixed description set. As for more elaborated approaches along these lines, cf. for example [10, 13].)

[3]The 'emotional domain' is most likely not the word but some constituent (noun phrases, etc.) or clauses. If we label on the word level we do not exclude any of these alternatives. In a subsequent step, we therefore can perform and assess several different types of chunking.

From a methodological point of view, our **7-class problem** is most interesting. However, the distribution of classes is very unequal. Therefore, we down-sampled *neutral* and *emphatic* and mapped *touchy* and *reprimanding*, together with *angry*, onto **A***ngry*[4] as representing different but closely related kinds of negative attitude. For this more balanced **4-class problem** 'AMEN', 1557 words for **A***ngry*, 1224 words for **M***otherese*, and 1645 words each for **E***mphatic* and for **N***eutral* are used, cf. [16]. Cases where less than three labellers agreed were omitted as well as those cases where other than these four main classes were labelled. We can see that there is a trade-off between 'interesting' and representative: our seven classes are more interesting, and our four classes are more representative, and therefore better suited for automatic classification, cf. [6].

4. NON-METRICAL MULTI-DIMENSIONAL SCALING

Input into Non-Metrical Multi-Dimensional Scaling (NMDS) is normally a matrix indicating relationships among a set of objects. The goal is a visual representation of the patterns of proximities (i.e., similarities or distances) among these objects. The diagonal (correspondence) is not taken into account; the matrices are either symmetric or are made symmetric, via averaging. The computation encompasses the following steps: with a random configuration of points, the distances between the points are calculated. The task is to find the optimal monotonic transformation of proximities (i.e., of the distances), in order to obtain optimally scaled data (disparities); the so-called stress-value between the optimally scaled data and the distances has to be optimized by finding a new configuration of points. This step is iterated until a criterion is met. The output of NMDS is an n-dimensional visual representation; one normally aims at two dimensions, one being often not interesting enough, and three often being difficult to interpret and/or not stable because of sparse data. The criteria for the goodness of the solution is the stress value and the RSQ value, together with interpretation quality – the last one admittedly being a rather vague but at the same time, very important criterion. The axes are meaningless, the orientation is arbitrary. We can interpret clusters and/or dimensions and, by that, we can find more general concepts than the single items (categories, labels) that were input into NMDS. Note that it is not the exact distance between items that should be interpreted and replicated but the basic configuration. Most useful is NMDS for exploration of new (types of) data. We will use the ALSCAL procedure from the statistical package SPSS.

4.1 NMDS solutions for our data

The MV cases described above we will call **absolute majority (AM)** cases; in addition, we define as **relative majority (RM)** those cases where a relative majority or no majority at all (i.e., equal distribution) is given. By that, we sort of **preemphasize** the non-MV cases.[5] Table 2 shows the number of cases per constellation, and Table 3 shows the combined confusion matrix for all labels, i.e., for AM and RM cases in percent.[6] To give two examples: For an AM case with a majority of 3/5 for **A**ngry, we enter 3 cases in the reference line into the cell for **A**ngry and the other two as 'confused with' into the cells for the resp. other labels in the same line. For a RM case with 1+1+1+1+1+1, i.e., equal distribution, we enter five times in turn each of the five different labels as reference and the other four as 'confused with' into the cells for the resp. other labels.

Table 2: Emotional labels used with # of majority voting MV

absolute majority AM	#
3/5	13671
4/5	17281
5/5	12751
relative majority RM	**#**
2+1+1+1	1554
2+2+1	3070
1+1+1+1+1	81
total	48408

Table 3: confusion matrix for AM and RM in percent

label	**A**	**T**	**R**	**J**	**M**	**E**	**N**
Angry	43.3	13.0	12.9	0.0	0.1	12.1	18.0
Touchy	0.5	42.9	11.6	0.0	0.9	13.6	23.5
Reprim.	3.7	15.6	45.7	0.0	1.2	14.0	18.1
Joyful	0.1	0.5	1.0	54.2	2.0	7.3	32.4
Mother.	0.0	0.7	1.4	0.8	61.0	4.8	30.3
Emphatic	1.3	5.7	6.7	0.5	1.2	53.6	29.8
Neutral	0.3	2.1	1.4	0.4	2.7	13.9	77.8

Fig. 1 shows the 2-dimensional NMDS solution for Table 3. As mentioned above, axes and orientation are arbitrary; in the following interpretation of our NMDS solutions given in Figures 1 to 5, the underlying dimensions are thus not identical with the axes ('Dimension 1', 'Dimension 2') displayed in the figures, and they are not necessarily orthogonal to each other. If we want to refer to the dimensions we interpret for our solutions, we will use the terms which refer to the compass rose: *west to east* thus means more or less along the x-axis, *south-west to north-east* means bottom left to upper right. Note that by that, we do not indicate any precise direction but only a rough orientation. *neutral* and *emphatic* cluster together, close to the origin. The first, most important dimension can be interpreted as VALENCE *west to east*. The second dimension is, however, not something like the 'traditional' dimension AROUSAL, but rather something that can be described as ORIENTATION

[4]If we refer to the resulting 4-class problem, the initial letter is given boldfaced and recte. Note that now, **A***ngry* can consist, for instance, of two *touchy* and one *reprimanding* label; thus the number of **A***ngry* cases is far higher than the sum of *touchy*, *reprimanding*, and *angry* MV cases.

[5]Preemphasis increases, for instance in audio signals, the magnitude of higher frequencies w.r.t. lower frequencies. If we 'preemphasise' our RM cases, we assign these rare cases higher weight by using the same case several times as reference. Another analogy is the logarithmic presentation of frequencies in a diagram if some classes have many tokens, some other only a few: here the bars for higher frequencies are lowered w.r.t. the bars for lower frequencies.

[6]In the tables, percent values per line sum up to 100%, modulo rounding errors. The labels are given recte, with boldfaced initials (row); for the columns, only the (unique) initials are given.

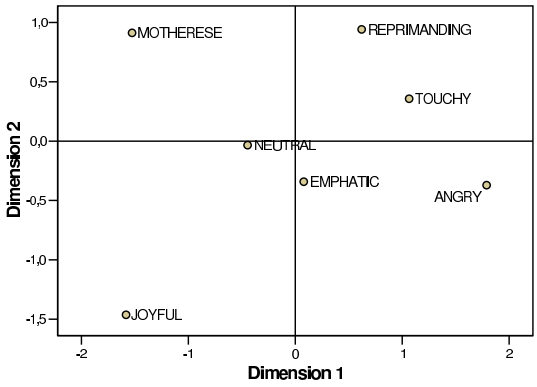

Figure 1: *NMDS solution for MV data with # > 50, 2 dimensions; stress: .23, RSQ = .82*

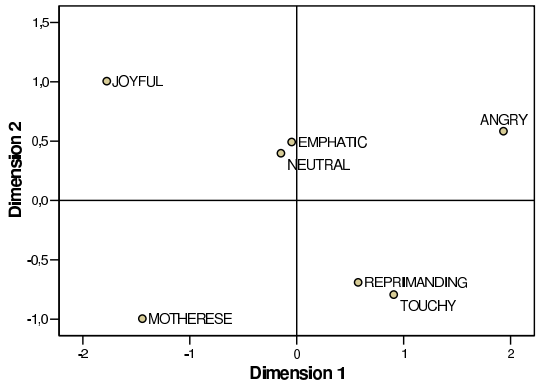

Figure 2: *NMDS solution for 'probability' data with # > 50, 2 dimensions; stress: .21, RSQ: .85*

(*south to north*) towards the subject him/herself or towards the partner (in this case, the AIBO), as DIALOGUE aspect (MONOLOGUE vs. DIALOGUE), or as [+/- INTERACTION]. In the following, we will use INTERACTION as term.[7] User states like *angry*, i.e., [- VALENCE], and *joyful*, i.e., [+ VALENCE], represent [- INTERACTION], whereas user states like *reprimanding*, i.e., [- VALENCE], and *motherese*, i.e., [+ VALENCE], represent [+ INTERACTION].

The computation of the confusion matrices might affect the dimensional solution. Thus for Table 4, another computation was chosen: each cell represents the probability for a word to be labelled with one emotion (line) by one labeller and with the same or another emotion (row) by another labeller, averaged across all 10 possible combinations of labellers: $P(A \leftrightarrow B)$; the values of all cells in the triangular matrix sum up to 100. This raw matrix, however, does not yield any meaningful dimensional solution because distribution in the cells is very unequal. Therefore, we normalized each line; by that, the values in percent of each line sum up to 100%. Thus for Table 3 we sort of 'preemphasised' the unclear, mixed cases, for Table 4 we sort of 'preemphasised' the rare cases.

Table 4: confusion matrix for 'probability' values in percent (cf. explanation in text)

label	**A**	**T**	**R**	**J**	**M**	**E**	**N**
Angry	15.4	16.7	12.8	0.1	0.1	17.6	36.7
Touchy	3.6	12.8	11.1	0.1	1.2	19.9	49.2
Repr.	3.4	14.1	17.8	0.2	2.2	24.5	37.1
Joyful	0.1	0.6	0.7	17.6	4.7	9.4	64.3
Mother.	0.0	0.9	1.2	0.7	32.8	5.8	58.1
Emphatic	0.7	3.5	3.4	0.3	1.5	21.2	68.7
Neutral	0.3	2.2	1.3	0.6	3.6	17.0	73.9

[7]Actually, the other names might be, in other contexts, even more adequate depending on the specific theoretical and empirical background: if communication is restricted to speech (for instance, via telephone), we might prefer dialogue vs. monologue (i.e., speaking aside). At least in German, verbs with this type of [+ INTERACTION] tend to be more transitive, i.e., having more valence slots than verbs with [- INTERACTION].

Fig. 2 displays the 2-dimensional solution for the matrix of Table 4. This is a nice demonstration that axes are meaningless and orientation is arbitrary but the general picture remains the same: *neutral* and *emphatic* cluster together close to the origin, *joyful* and *motherese* are positive, i.e., [+ VALENCE] and [-/+ INTERACTION], *angry* is like *joyful* but negative, i.e., [- VALENCE]. *Reprimanding* is close to *touchy* both in Figures 1 and 2; however, the dimensional solution for the MV data in Fig. 1 can be interpreted easier than the one for the probability data in Fig. 2 because *reprimanding* is more closely attached to [+ INTERACTION] than *touchy* - an argument in favour of this type of computation of MV?

As mentioned in section 3, for automatic classification, cf. [16, 6], we mapped our labels onto a 4-class problem. Table 5 displays the confusion matrix for these four labels, computed the same way as Table 3. In Fig. 3, the 2-dimensional NMDS solution for the confusion matrix of Table 5 is shown. There are only four items, this 2-dimensional solution is therefore not stable. **N***eutral* is close to the origin, as expected; the first dimension seems to be VALENCE (*west to east*) again: from **M***otherese* to **N***eutral* to **E***mphatic* to **A***ngry*. However, the second dimension is not easy to interpret.

As usual in research on realistic emotions, we are facing a sparse data problem: with less representative data, we can find interesting dimensions but of course, automatic classification performance is not high, cf. [6]. With (statistically) representative data – obtained via mapping onto cover classes – classification performance is higher but our interesting dimension INTERACTION is gone, i.e., no longer visible.

Table 5: confusion matrix for AMEN

label	**A**	**M**	**E**	**N**
Angry	70.6	0.4	10.7	18.2
Motherese	0.4	68.8	1.5	29.3
Emphatic	5.7	0.2	65.5	28.5
Neutral	2.1	2.6	13.3	82.0

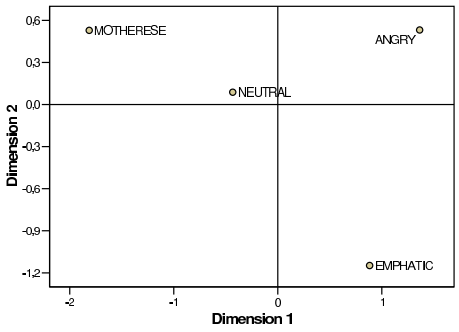

Figure 3: *NMDS solution for the 4-class problem AMEN, 2 dimensions; stress: .19, RSQ: .90*

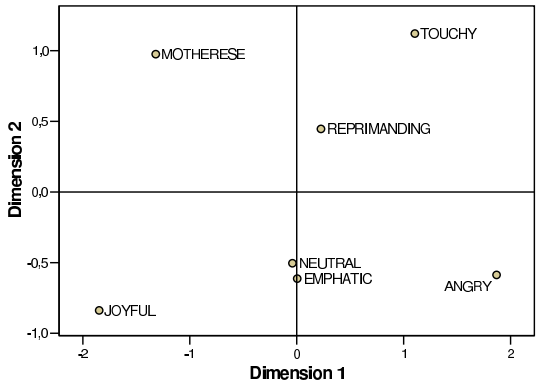

Figure 4: *NMDS solution for English MV data, 2 dimensions; stress: .17, RSQ: .89*

4.2 Other Types of Data

If data are not pre-defined, i.e., if we only can label what we can find in realistic databases, then we will most likely find something different – even different categories and by that, different dimensions – for different types of databases. To illustrate this aspect, we first computed a 2-dimensional NMDS solution for our parallel English data, exactly along the same lines as for our German data: MV, 'preemphasis'. The English data do not only represent another language but differ in several aspects slightly from our German data: there were 30 English children which took part, with a wider range of age, namely between 4 and 14. There were two recordings, the second being parallel to one of our subdesigns (so called 'parcours'; details can be found in [3]). In the first recording, the same sub-design was used but the AIBO behaved obedient and followed the children's commands. The children were not told that they could communicate with the AIBO as with a friend. The data was annotated by three out of the five labellers which annotated our German data. MV therefore means that two out of three labellers agreed. This is a typical situation that we often face in daily practice: parallel does not mean strictly parallel – for our English data, there are, e.g., less subjects, age distribution is different, there are less labels and less labellers. Fig. 4 displays the 2-dimensional NMDS solution for the English data. For comparison, we take exactly the same labels as we did for our German data, even if MV frequency is now sometimes below 50, cf. Table 1. We can find our two dimensions, we can replicate the clustering found in Figures 1 and 2; what we do not find is the exact position of some of our categories in this space: *touchy* and *reprimanding* changed places (note that there are only seven *touchy* cases), and *neutral* and *emphatic* are not close to the origin. If we consider that the sparse data problem for our English data is even more pronounced than for our German data, cf. Table 1, this is a reassuring result.

But now we now want to have a look at the dimensions we can extract for data obtained within a totally different scenario, namely a call-center scenario: the German SympaFly database was recorded using a fully automatic speech dialogue telephone system for flight reservation and booking. In the first stage of this system, performance was rather poor (approx. 30% dialogue success rate); in the last, third stage, performance was very good (above 90% dialogue success rate). Recordings were made with volunteering subjects (2. stage) and with employees of a usability lab (1. and 3. stage). A full description of the system and these recordings can be found in [4, 5]. We employed two labellers; as is the case for the AIBO labels, the labels were chosen in a pilot pass. The confusion matrix, this time with the absolute number of items in each cell in order to indicate the sparse data problem more clearly, is given in Table 6. Note that here, we annotated whole turns and not words. Each turn had 4.3 words on average.

Fig. 5 shows for those items with a frequency above 50 for each of the two labellers the 2-dimensional solution for the SympaFly data. With only two labellers, there is no MV. We therefore took each labeller in turn as reference (line), normalized each line summing up to 100%, and computed the mean percent value per cell for these two matrices. (Needless to say that this solution can only be taken as some indication because we only have two labellers, and because the distribution of our items is extremely unequal.) It is self-evident why we do not find the INTERACTION dimension that is specific for our AIBO data: call center clients do not use *motherese* or this specific type of *reprimanding* while communicating with a human operator, let alone with an automatic system. However, again we do not find the clear-cut dimensions AROUSAL or VALENCE. The first dimension (*south-west to north-east*) might be another type of INTERACTION (related to CONTROL): the normal one in the case of *neutral* and *emphatic*, and withdrawal from normal interaction, i.e., rather some sort of meta-communication, in the case of *helpless* and *ironic*. The second dimension (*south-east to north-west*) could be some sort of EXPRESSIVITY – related to but not necessarily identical with AROUSAL: it is typical for *ironic* that it lacks EXPRESSIVITY the same way as *neutral* does – otherwise, it would no longer be irony. *touchy* on the other hand, displays EXPRESSIVITY.

The chunking of *neutral* and *emphatic* can be observed throughout in all figures and is consistent with our explanation in section 3 that *emphatic* does not necessarily indicate any (strong) deviation from a neutral state.

5. DISCUSSION

In this section, we want to discuss several aspects and questions in more detail.

L1 ↓ | L2 →	J	N	S	I	C	E	A	P	H	T	Total
Joyful	12	5	-	3	-	-	-	-	-	-	20
Neutral	13	5355	3	31	18	110	1	6	31	72	5640
Surprised	-	1	3	1	-	1	-	-	1	-	7
Ironic	4	17	1	28	1	1	-	-	2	8	62
Compassionate	-	-	-	-	-	-	-	-	-	-	
Emphatic	2	340	-	8	11	218	2	8	7	54	650
Angry	-	2	-	-	-	-	-	-	2	4	8
Panic	-	1	-	-	-	-	-	7	-	-	8
Helpless	-	16	-	5	2	1	-	2	21	9	56
Touchy	2	39	-	1	-	21	1	-	3	76	143
Total	33	5776	7	77	32	352	4	23	67	223	6594

Table 6: SympaFly: Confusion matrix for emotional user states annotated per turn, two labellers

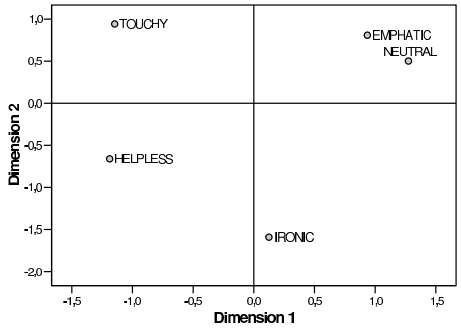

Figure 5: *NMDS solution for SympaFly (call-center data) with # > 50; stress: .24, RSQ: .80*

5.1 Assessment of NMDS Solutions

The rule of thumb is that stress values below .2 and RSQ values above .8 are OK. Note that this should be taken only as a rough guide: it strongly depends on the type of data. Studies cannot be compared in a strict sense; however, it is plausible that more artifical and by that, more controlled data will, other things being equal, result in a better quality. For instance, acted facial expressions in [14] yielded better stress and RSQ values, and resp. values are very good in [11] even in a 1-dimensional solution for smilies which of course do have very unequivocal characteristic traits. In contrast, we can expect much more 'white noise' in our realistic data especially if the emotional states are not full-blown but mixed. In [6] we show that there obviously are more or less clear cases: the better performance of **prototypes** in automatic classification indicates that the emotional user states labelled are either a graded or a mixed phenomenon – or both.[8]

There might be some 'critical mass' w.r.t. number of items in a NMDS, and number of different labellers: if the number of items is too small w.r.t. the dimensionality, then the solution is not stable. If the number of labellers is too small, then spurious and random factors might influence computation. The one and/or the other factor might be responsible for the constellations in Figures 3 and 5. However, it is reassuring that different computations yield similar solutions in the case of Figures 1, 2 and 4.

[8] At the moment, we therefore run an additional annotation pass where the labellers can annotate more than one label for the same word and assign percentages for each label summing up to 100%. Note that due to sparse data, this can only be done for the four AMEN classes.

5.2 Mixed Cases

In Table 7 we give two interesting examples of a relative majority for mixed cases; in the left row, the German words belonging to one utterance are given; non-standard forms such as *ne* instead of *nein*, are starred. In the right row, the English translation is given. In between, the labels given by labeller one (L1) to five (L5) are displayed. We can see that in the first example, *motherese* alternates with *reprimanding* (and *neutral*). Thus, INTERACTION is clearly positive, although VALENCE is not that clear. Obviously, if *motherese* is labelled, the 'tone of voice' was the discriminating feature, if *reprimanding* was labelled, the semantics of *'no'* played a greater role. In the second example, the negative VALENCE is clear, the detailed classes obviously are not. A mapping onto a cover class *negative* or **A***ngry* thus suggests itself, cf. as well the similarities of these negative labels in Table 3. However, we have to keep in mind that only some few cases display such mixed annotations, cf. above Table 3. The cases are thus 'interesting', but – at least for our data – not necessarily representative. By using preemphasis, we do account for such mixed cases in our NMDS solutions as well.

5.3 Different Conceptualizations

Figure 6 shows for our 4-class AMEN problem a scatterplot with the distribution of **M***otherese* vs. **A***ngry* per speaker (leaving aside one outlier subject which displays very high frequencies for both). Spearman's rho (non-parametric correlation) for these two distributions is .47 (without the outlier) or .50 (with the outlier). There seem to be, however, two distinct trends in this plot: one type of children tends towards using **A***ngry* but not (much) **M***otherese*, another type uses both. Maybe we can even tell apart three different interaction types: one addresses the robot as a sort of remote control tool, without showing much emotions. The second one is sort of mixed, showing anger sometimes, and the third one addresses the AIBO really as an interaction partner, as a real pet: encouraging, if need be, and reprimanding, if need be. Here, the **target prototypes** are thus at the origin (no interactive behaviour at all, only commands), high on the y-axis and low on the x-axis (showing only **A***ngry*), and high on both axes (showing both **M***otherese* and **A***ngry* which means a fully developed interactive behaviour). If

Table 7: Examples for Relative Majority = 2

German	L1	L2	L3	L4	L5	English
mixed VALENCE, clear INTERACTION						
**ne*	M	R	N	M	R	*no*
**ne*	M	R	N	M	R	*no*
**ne*	M	R	N	M	R	*no*
so	M	R	N	M	N	*so*
weit	M	R	N	M	N	*far*
**simma*	M	R	N	M	N	*we are*
noch	M	R	N	M	N	*yet*
nicht	M	R	N	M	N	*not*
aufstehen	M	R	N	N	R	*get up*
clear VALENCE, unclear categories						
nach	A	T	E	E	N	*to*
links	A	T	E	E	R	*the left*
Aibo	A	T	T	R	R	*Aibo*
nach	A	T	T	E	N	*to*
links	A	T	T	E	R	*the left*
Aibolein	A	T	E	A	R	*little Aibo*
ganz	A	T	E	A	R	*very*
böser	A	T	T	A	N	*bad*
Hund	A	T	T	A	N	*dog*

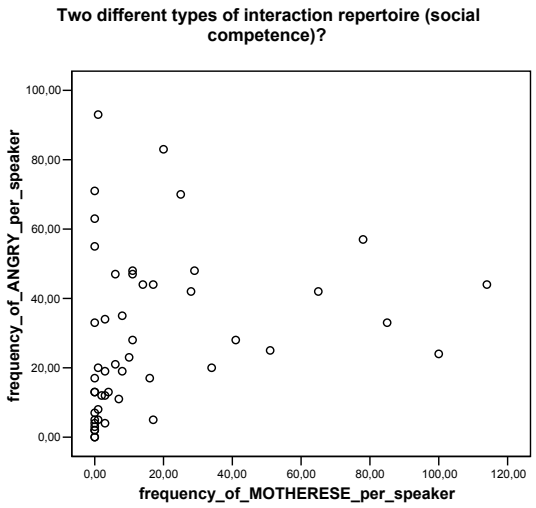

Figure 6: *Scatterplot: Distribution of Motherese and Angry per Speaker; displayed is # of cases*

children belong to the third type, we can conclude that they use a more elaborated linguistic and by that, interaction repertoire. It is an interesting question whether such an elaborated repertoire goes along with a higher social competence. Furthermore we can find out whether there are gender-specific differences: in our database, girls tend to use more **M***otherese* and less **A***ngry* than boys. This difference is, in a two-tailed t-test, not significant but in a one-tailed; as this difference was not formulated as alternative hypothesis, we had to use the two-tailed test.

It is clear that these different conceptualizations lead to different or missing dimensions: if subjects do not use **M***otherese* then NMDS will not find our second dimension INTERACTION. And if subjects neither use **M***otherese* or **A***ngry* (i.e., *touchy*, *reprimanding*, or *angry*), then we possibly will not find our first dimension VALENCE either.

5.4 How to Annotate, how to Process

There are indications that emotion-related user states (encompassing the states that we could find in our data) are more or less continuous. This does not tell us the best way how to annotate these phenomena, and it does not tell us either whether we will process them in an automatic system as dimensional entities or not. It has been our experience in fully developed end-to-end systems, cf. the SmartKom system [7, 15], that the highly complex processing makes it necessary to map any fine-grained scale onto some very few states - two or three. Early/late mapping and/or fusion can be imagined. It might be a matter of practicability and not of theoretical considerations whether we want to use categorical or graded labels as input into such systems. Moreover, if we go over to large-scaled collections of realistic databases, it might not be feasible to employ several labellers using a very elaborated annotation system.

5.5 Which Dimensions

The dimensions that best model specific types of scenarios depend crucially on at least: firstly, the subjects and their conceptualizations; secondly, the communication structure, e.g., whether it is symmetric or not; thirdly, in which setting the emotions are observed. Due to the observer's paradox [12, 4], the threshold for displaying emotions might be higher, the more likely it is that the subjects are being observed by a third party, meaning that some type of general public is present.

It might as well be that for some data, no clear picture emerges. This can be due to insufficient size of the database, or simply to a constellation where no clear dimensional solution can emerge. The dimensions we can find will definitely be affected by the sparse data problem: for our SympaFly data we decided not to take into account labels with a frequency below 50 in order to ensure a half-decent robustness of our solution. By that, we excluded user states like *angry* and *panic* from our analysis; with these emotions, we probably could have obtained AROUSAL as first or second dimension. Thus what we get is an indication of those emotional user states we will encounter in applications if – and only if – the distribution of our phenomena and by that, labels, can be transferred to real applications. Of course, we cannot say anything about the emotions our subjects will display in other situations.

It will certainly not be meaningful to create a new dimensional space each time we deal with a new scenario. As far as we can see, it might often be the case that only a certain **sub-space** can be modelled with those categories that can be found and labelled in specific databases. On the other hand, even if it might be possible to map any new category onto the traditional dimensions AROUSAL and VALENCE etc. this will not be a very wise strategy because in many cases, this solution will not turn out to be stable and adequate.

6. CONCLUDING REMARKS

We might not exactly be on the verge of a classic paradigm shift but we definitely are mid stream: turning from theoretical playgrounds towards demands put forth by applications. In this situation, we favour a rather data-driven, 'roving' approach such as the one described in this paper, i.e., realistic, non-acted data and non-pre-defined sets of labels. Even if possibly, new models combining emotion with the interaction aspect might be grounded in such studies, our more modest goal is for the moment simply to get at a clearer picture of the data we will have to deal with in possible applications: a characterisation in terms of some few dimensions might be more informative than just using a list of categorical labels.

In conclusion and coming back to the title of this paper 'private emotions vs. social interaction': emotions are to a large extent rather private and therefore, we might not be able to observe them as often, esp. in 'public' settings. Instead, it might be necessary to model social interaction in more detail.

7. ACKNOWLEDGMENTS

This work was partly funded by the EU in the framework of the two projects PF-STAR under grant IST-2001-37599 and HUMAINE under grant IST-2002-507422, and by the German Federal Ministry of Education and Research (*BMBF*) in the framework of the two projects SmartKom (Grant 01 IL 905 K7) and SmartWeb (Grant 01IMD01F). The responsibility for the contents of this study lies with the authors.

8. REFERENCES

[1] E. André, L. Dybkiaer, W. Minker, and P. Heisterkamp, editors. *Affective Dialogue Systems, Proc. of a Tutorial and Research Workshop*, volume 3068 of *Lecture Notes in Artificial Intelligence*, Berlin, 2004. Springer-Verlag.

[2] A. Batliner, K. Fischer, R. Huber, J. Spilker, and E. Nöth. How to Find Trouble in Communication. *Speech Communication*, 40:117–143, 2003.

[3] A. Batliner, C. Hacker, S. Steidl, E. Nöth, S. D'Arcy, M. Russell, and M. Wong. "You stupid tin box" - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Proc. LREC 2004*, pages 171–174, Lisbon, 2004.

[4] A. Batliner, C. Hacker, S. Steidl, E. Nöth, and J. Haas. User States, User Strategies, and System Performance: How to Match the One with the Other. In *Proc. of an ISCA Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, pages 5–10, Chateau d'Oex, 2003.

[5] A. Batliner, C. Hacker, S. Steidl, E. Nöth, and J. Haas. From Emotion to Interaction: Lessons from Real Human-Machine-Dialogues. In André et al. [1], pages 1–12.

[6] A. Batliner, S. Steidl, C. Hacker, E. Nöth, and H. Niemann. Tales of Tuning – Prototyping for Automatic Classification of Emotional User States. Submitted to Interspeech 2005.

[7] A. Batliner, V. Zeissler, C. Frank, J. Adelhardt, R. P. Shi, and E. Nöth. We are not amused - but how do you know? User states in a multi-modal dialogue system. In *Proc. Eurospeech*, volume 1, pages 733–736, Geneva, September 2003.

[8] R. Cowie and R. Cornelius. Describing the emotional states that are expressed in speech. *Speech Communication*, 40:5–32, 2003.

[9] R. Cowie and M. Schröder. Piecing Together the Emotion Jigsaw. In S. Bengio and H. Bourlard, editors, *Proc. Machine Learning for Multimodal Interaction, First International Workshop, MLMI 2004, Martigny, 2004*, volume 3361 of *Lecture Notes in Computer Science*, pages 305 – 317. Springer, 2004.

[10] R. Craggs and M. M. Wood. A categorical annotation scheme for emotion in the linguistic content of dialogue. In André et al. [1], pages 89–100.

[11] R. Jäger and J. Bortz. Rating scales with smilies as symbolic labels – determined and checked by methods of Psychophysics. In *70. Annual Meeting of the International Society for Psychophysics*, Leipzig, 2001.

[12] W. Labov. The Study of Language in its Social Context. *Studium Generale*, 3:30–87, 1970.

[13] K. Laskowski and S. Burger. Development of an Annotation Scheme for Emotionally Relevant Behavior in Multiparty Meeting Speech. Submitted to Interspeech 2005.

[14] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba. Coding Facial Expressions with GaborWavelets. In *3rd International Conference on Face & Gesture Recognition (FG '98), April 14-16, 1998, Nara, Japan*, pages 200–205. IEEE Computer Society, 1998.

[15] T. Portele. Interaction Modeling in the SmartKom system. In André et al. [1], pages 89–94.

[16] S. Steidl, M. Levit, A. Batliner, E. Nöth, and H. Niemann. "Of All Things the Measure is Man": Automatic Classification of Emotions and Inter-Labeler Consistency. In *Proc. ICASSP 2005, Philadelphia*, 2005.

[17] W. Wundt. *Grundzüge der Physiologischen Psychologie*, volume 2. Engelmann, Leipzig, 1903. original published 1874.